Common Data Elements (CDEs) Harmonization

August 24, 2004
Tommie Curtis
SAIC, Contractor for NCICB

Agenda

- Harmonization Goals
- Use of National/International Standards
- Harmonization Tools – caDSr, EVS
- Metadata Consistency
- Registration/Workflow Status
- Harmonization Candidates
- Harmonization Process
- Questions and Answers

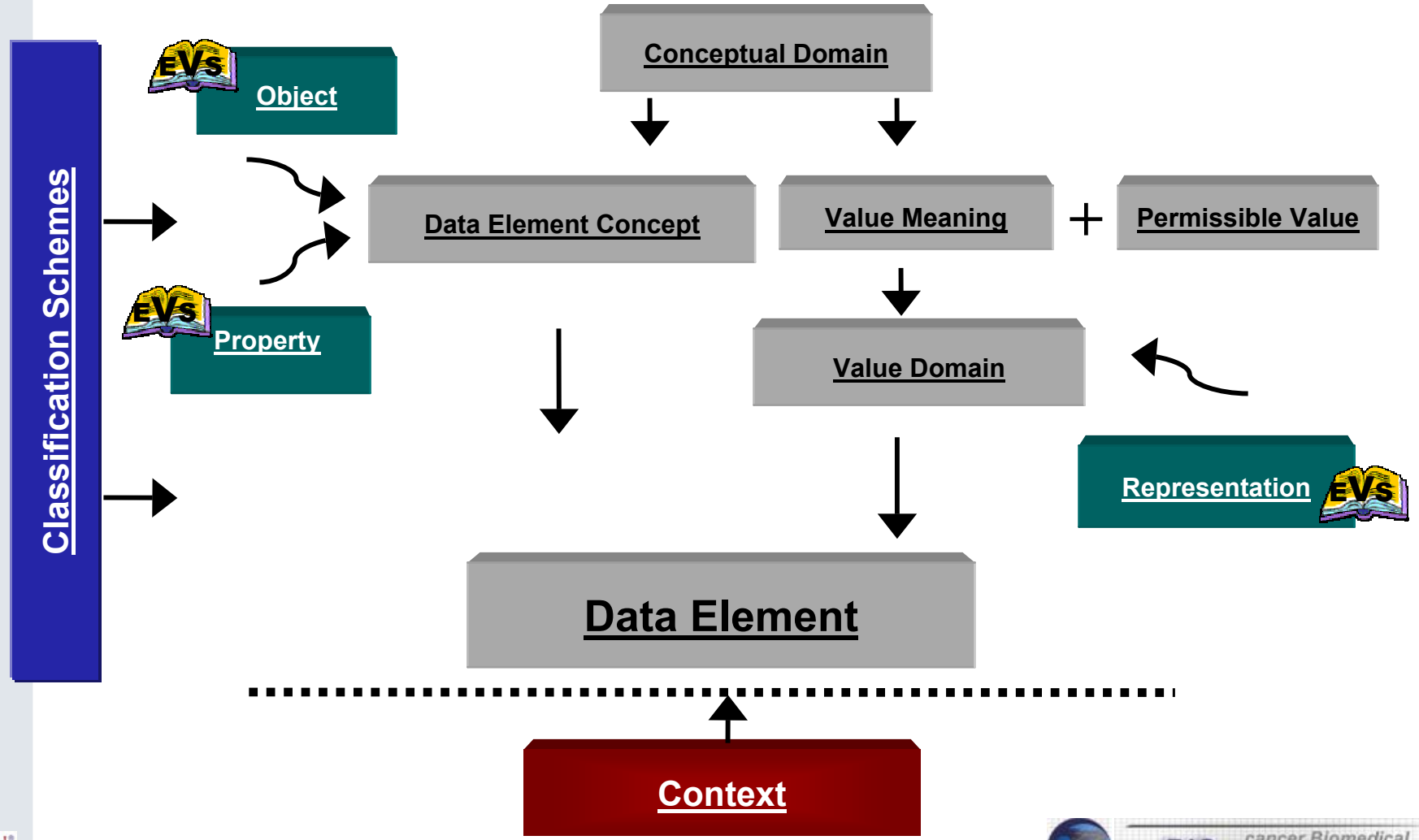
CDE Harmonization Goals

- Use a unified, shared set of best practices for CDE development.
- Survey and examine external metadata projects in life science informatics and identify a strategy to become compatible with those efforts that are deemed important to caBIG's mission.
- Adopt an appropriate use of terminology from the EVS for CDE development.
- Facilitate consensus among participating caBIG workspaces administrators to identify and implement a mechanism to create and maintain standard, ISO/IEC 11179 compliant, nonredundant CDEs across all the caDSR contexts.
- Establish a common conceptual classification system to organize the CDEs across workspaces.

Adopt and Harmonize with National and International Standards

- Identify health information standards for developing CDEs and value domains. Provide standard CDEs and value domains supported by other organizations, e.g., Health Level Seven (HL7), MicroArray and Gene Expression (MAGE), and regulatory bodies, such as the Food and Drug Administration (FDA).
- Identify other data standards, (e.g., International Organization for Standardization [ISO], Federal Information Processing Standards [FIPS]) that relate to the caDSR CDEs.
- Develop a plan, where needed, for harmonization with external standards.
- Move toward using data that will be compatible with other governmental efforts in development and use of common terminology.
- Provide a plan for presenting the standards information in the caDSR to enhance discovery and reuse.

ISO/IEC 11179 Information technology – Metadata registries, Part 3: Registry metamodel and basic attributes



caDSR Role

- Location of workspace CDE metadata.
- Source of CDEs for reuse.
- Tool for harmonization review.
- Provides a link to terms and concepts.

EVS Role

- Location of terminology and definitions.
- Source of terms for reuse.
- Tool for harmonization of terms and concepts.
- Linked to caDSR and models.

Harmonization Agreements

- For CDEs to be shared there must be agreement on the data element definitions, value domains, and permissible values.
- Ongoing cross-context designation provides a basis for harmonization efforts.
- caBIG context will own and manage classification schemes and data elements for harmonized CDEs.
- Registration status will be adopted to indicate consensus progression.

Metadata Consistency

- Use a registration process that will minimize metadata content variability and enhance reuse.
- Adopt a “business rules framework” for populating the caDSR so that information is entered in a common way across contexts.
- Use a common vocabulary across contexts.
- Develop a work plan for applying the cross-context business rules.
- Maintain the credentials of the caDSR content.

Using Registration Status to Convey Approval of CDEs

1. Standard - the Administered Component has all necessary metadata, has met all quality requirements, and has been approved for usage in all contexts.
2. Candidate - the Administered Component has all the necessary metadata, has met all quality requirements, and has been proposed for usage in all contexts.
3. Qualified - the Administered Component has all the necessary metadata, has met all quality requirements, and is ready to be reviewed for usage in all contexts.
4. Standardized Elsewhere - the Administered Component has been adopted for standard usage by a community outside of NCI.
5. Superceded - the Administered Component is no longer recommended for use in the NCI community because it has been replaced by another preferred Administered Component.
6. Retired - the Administered Component is no longer recommended for use as a NCI standard.
7. Suspended - the Administered Component has been considered for standard status in the past and was not selected for promotion.
8. Application - the Administered Component is part of a computer program or application and may not have all the metadata generally required by the registry.

Using Workflow Status to Convey Progress of CDEs in a Review Process

1. Released - the Administered Component has been fully reviewed and approved by the owning Context's review process. Administered Components with this workflow status may be used by all contexts.
2. Approved for Trial Use - the Administered Component is included on CDE-compliant forms.
3. Committee Approved - the Administered Component has been adopted for use by a disease committee.
4. Committee Submitted Used - the Administered Component can be used on a trial but has not been fully reviewed and approved.
5. *Draft New* - the Administered Component is being developed and may not be fully specified or reviewed.
6. *Draft Mod* - the Released version of an Administered Component has been changed and is in the review and approval process.
7. *Retired Withdrawn, Retired Phased Out, Retired Archived, Retired Deleted* - the status assigned to Administered Component no longer available for use.
8. Released Non-compliant - the Administered Component may be used but is not fully specified or approved.

Administered Components with Released, Approved for Trial Use, or Committee Approved should be considered first for reuse.

Administered Components with *Draft New, Draft Mod,* or any of the *Retired* statuses should NOT be considered for reuse.

Harmonization Candidates Based on Commonly Used Data

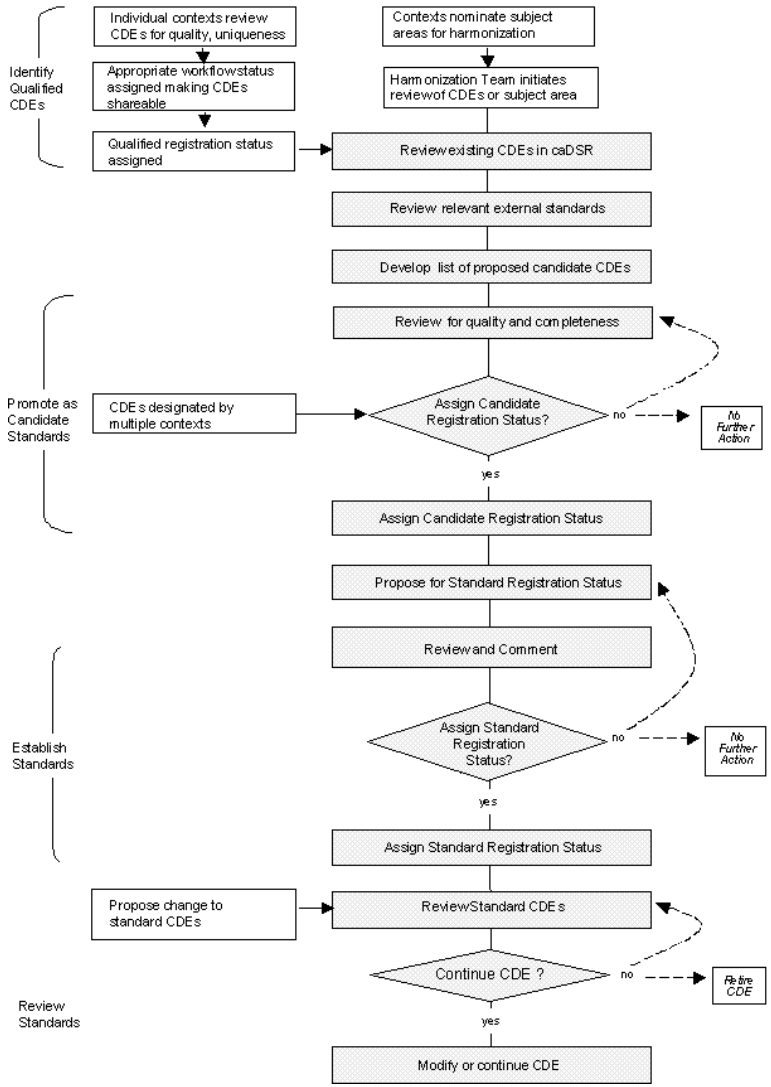
The Context Administrators suggested these areas:

- ✓ Patient Demographics (including Patient Identification, Patient Medical History, Eligibility, Enrollment)
- ✓ Gene Identification, Description, and Expression Information
- ✓ Contact Information (including Name, Organization, Address)
- ✓ Study Related Information (including Study Identification, Study Management, Study Descriptive)
- ✓ Laboratory Results
- ✓ Study Drug Description
- ✓ Serious Adverse Event/Adverse Event Description

Harmonization Process

- Workgroups Propose CDEs for Use.
- Workspace Review and Recommendation.
- Vocabulary/CDE Team Review.
- Assignment of Candidate Registration Status.
- Propose for Standard Registration Status.
- Confer Standard Registration Status.
- Use Across Workspaces.
- Ongoing Maintenance.

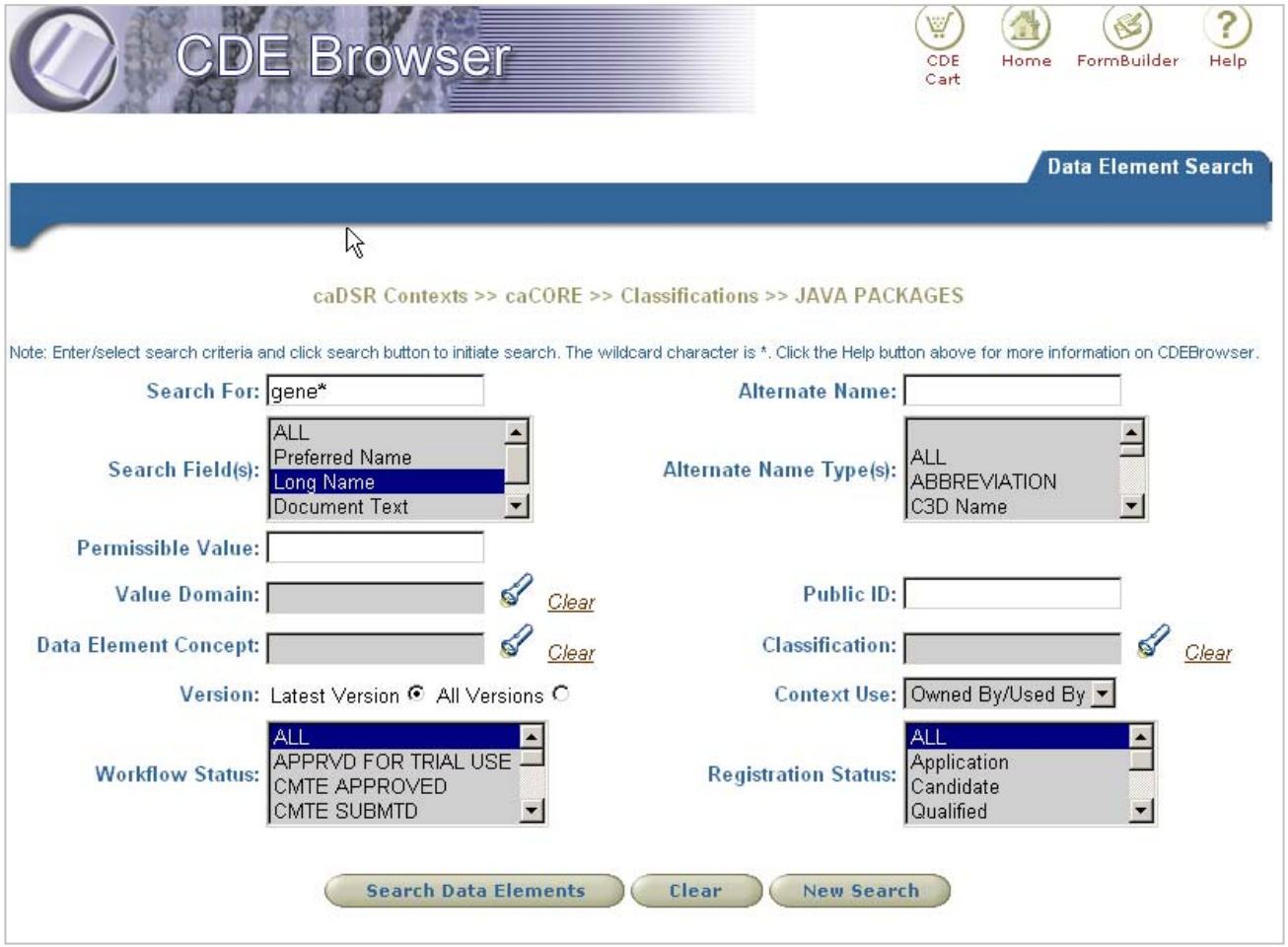
CDE Harmonization Process



Recommended Readings

- ISO 11179 Part-1, Framework and specifications, Section Appendix A.2, Classifications.
- ISO 11179 Part-2, Classification of data elements.
- ISO 11179 Part-3, Basic attributes of standardized data elements.
- ISO 11179 Part-4, Rules and guidelines for formation of definitions.
- ISO 11179 Part-5, Naming and identification principals for data elements.
- ISO 11179 Part-6, Registration of data elements.
- ISO 20944-02 Metadata registry interoperability and bindings – Part 2: Attribute mapping, Section 4, Identifier mangling conventions.
- ISO 20943- Part 1: Metadata registry content consistency – Part 1: Data Elements. Section 6: Bottoms-up approach Section 7: Top-down approach.
- ISO 20943- Part 3: Metadata registry content consistency – Part 1: Value Domains.
- Common Data Element Harmonization Tactical Action Plan. ftp://ftp1.nci.nih.gov/pub/cacore/caDSR/CDE_Harmonization_TAP.doc
- caDSR documentation and business rules http://ncicb.nci.nih.gov/core/caDSR.

Example Search for “gene”



Browser Search Results

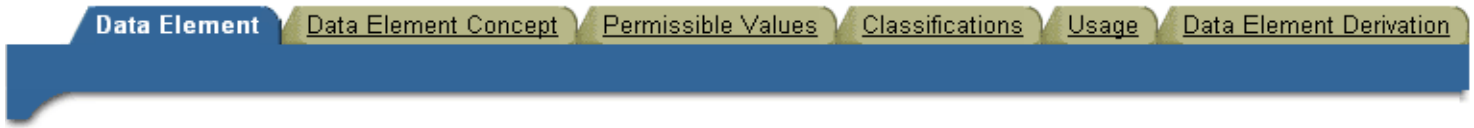
Add to CDE Cart

1 - 4 of 4

Long Name	Document Text	Owned By	Used By Context	Registration Status	Workflow Status	Public ID	Version
geneInfold		caCORE			DRAFT MOD	2178609	2.1
geneInfold		caCORE			DRAFT MOD	2178635	2.1
generalTerm		caCORE			DRAFT NEW	2184511	2.1
geneticManipulation		caCORE			DRAFT MOD	2181064	2.1

1 - 4 of 4

Result 1



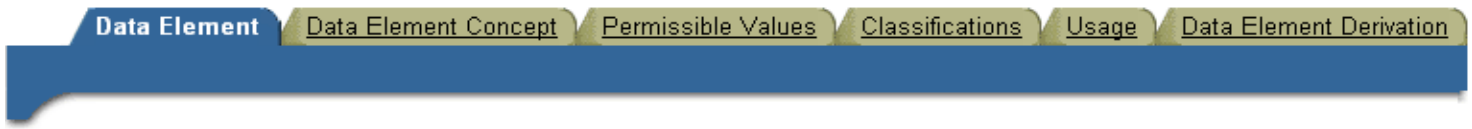
Data Element Details

Public ID:	2178609
Preferred Name:	ProteinGeneInfold
Long Name:	geneInfold
Document Text:	
Definition:	The gene information identification for the protein object.
Value Domain:	java.lang.String
Data Element Concept:	Protein
Context:	caCORE
Workflow Status:	DRAFT MOD
Version:	2.1
Origin:	
Registration Status:	

Reference Documents

Document Name	Document Type	Document Text	URL
caBIO-2.1.mdl (May2004)	REFERENCE	caBIO Domain Object Model	
caBIO-2.1.mdl (May2004)	LABEL	ProteinGeneInfold	
caBIO-2.1.mdl (May2004)	UML Attribute	geneInfold	

Result 2



Data Element Details

Public ID:	2178635
Preferred Name:	ProteinHomologGeneInfold
Long Name:	geneInfold
Document Text:	
Definition:	The gene identifier of the gene associated with the protein.
Value Domain:	java.lang.String
Data Element Concept:	ProteinHomolog
Context:	caCORE
Workflow Status:	DRAFT MOD
Version:	2.1
Origin:	
Registration Status:	

Reference Documents

Document Name	Document Type	Document Text	URL
caBIO-2.1.mdl (May2004)	REFERENCE	caBIO Domain Object Model	
caBIO-2.1.mdl (May2004)	LABEL	ProteinHomologGeneInfold	
caBIO-2.1.mdl (May2004)	UML Attribute	geneInfold	

Questions and Answers